社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2008－MPS－72　(15)
2008－BIO－15　(15)
2008／12／17

ラフ集合を用いた分類とその適用

加島　智子 1，Sophia Lin2，石井　博昭 1，和多田　淳三 2
1 大阪大学大学院　情報科学研究科
2 早稲田大学大学院　情報生産システム研究科

本論文ではラフ集合によってデータの分類を行っている．細分された部分集合がより大きな決定集合に分類することがでるかを識別する．ラフ集合の概念によって，データを自動的に分類するモデルを提案している．ここではデータ分類を効率的に実現することを目的とし，高齢化問題に本手法を応用している．近年，高齢化社会の問題に直面し，解決を模索している．そこで本論文では識別問題として，様々な都市が高齢化社会に分類されるか否かを議論している．

A Rough Set Approach to Classification and Its Application

Tomoko Kashima[1], Sophia Lin[2], Hiroaki Ishii[1], Junzo Watada[2]
[1] Graduate School of Engineering, Osaka University
[2] Graduate School of Information, Production and Systems, Waseda University

The objective of this paper is to realize a simple classification method in a rough set approach that distinguishes whether a subset can be classified in the target set or not. The algorithms of Rough Set will be used to analyze the data and in order to illustrate the method, we just use some artificial data in this paper. As its application, we discuss the aged society that should influences on policy making. The problem of aged society has become more and more severe all over the world. Almost all the countries have to face and solve this problem. In this problem the distinguish is done whether cities can be classified into an aged one or not.

1　はじめに

本論文ではラフ集合を用いてデータ分類を行う[1]．ラフ集合は有効なデータ分類手段であり，注目を集めている．そこでラフ集合を用いて国または都市が高齢化社会に含まれるか否かラフ集合によって決定する方法を検討する．

ラフ集合は，1980 年代に Z.Pawlak によって数学的なアプローチとして，データを分析するために開発された．新しいデータ分析手法として，ラフ集合はすべての種類の不確実なデータ，矛盾したデータおよび不完全なデータに有効的である．また，ラフ集合は情報システムに有効な知識獲得すことができる．このように他のデータ分類アプローチと比較して，ラフ集合は多くの長所がある．1980 年代以来，ラフ集合はその数学的理論，アルゴリズム，それらを実問題に広く応用してきた．

特に現在は，ラフ集合の数学的理論とアルゴリズムが注目を集めている．数学的な理論の研

究は，演算子の構造，ラフ集合の空間，ラフ集合理論の拡張などがある．また，アルゴリズムに注目する研究には，ラフ集合理論の縮小アルゴリズム，決定ルール抽出のアルゴリズムなどがある[2-4]．

2 ラフ集合理論

ラフ集合は 1982 年に提唱された理論で，感性工学の分野で用いられてきた．我々が何か対象を識別しようとするとき，粗い記述は対象を十分に特定できないというデメリットがあり，一方，細かい記述は対象をより精密に特定するものの，本質が見極めにくくなりやすいという欠点を持っている．ラフ集合は対象の集合をうまく特定できる範囲で情報を粗くすることで，対象の集合の程よい記述を求めることが可能である．

3. モデル

本論文ではラフ集合のアルゴリズムにより，データを自動的に分類するモデルを提案している．このモデルではデータ分類の効率化を実現することを目的としている．つまり，もし多くのデータを持っているならば，いくつかの特徴によりそれらを分類することが可能となる．データを入力するだけで自動的に分類される．この方法により，特定のグループに含まれるか否か簡単に判別することが可能となる．

図 1 はラフ集合の定義を示している．図 1 の集合 X の内側部分は正領域，集合 1 から集合 17 までの集合 X に重なる部分は境界域であることを表している．図のように正領域は集合 X に完全に含まれている．しかし，集合 16 と集合 17 は集合 X に大部分が含まれており，また集合 1 と集合 2 は集合 X にあまり含まれていない．したがって，集合 1 から集合 17 のような境界域が集合 X に含まれているかどうか決定する方法を考えていく．方法として，各集合に対してしきい値を持たせる．集合の要素がしきい値を越える場合，部分集合に含むまたは，含まないとする．

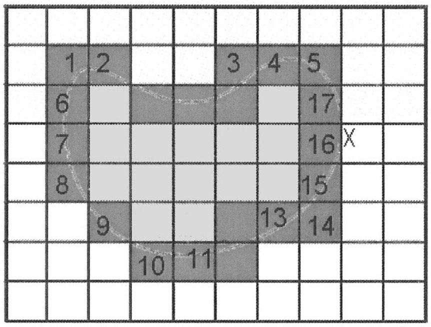

図 1 ：ラフ集合のイメージ

上記の説明に従い，本論文では，モデルのアルゴリズム設計を行った．全体集合に含まれる 30 の部分集合があり，それぞれの部分集合は 50 の標本値を持っていると仮定する．さらに，全体集合の判別比があり，それは各部分集合の分類結果が判別比より大きい場合，その部分集合は集合 X に含むことができるとする．このモデルを用いることにより，分類問題の結果を容易にかつ迅速に得ることが可能となる．このモデルの設計手順はステップ 3.1，3.2．3.2 で述べる．

3.1 標準的な確率変数に対する値の生成

まず，各部分集合は正規分布に従うと仮定し，現実の状況と良く似たサンプルデータを作成する．次に集合 A は 30 の部分集合を持ち，各部分集合は 50 の標本値を持つとする．値はボックス＝ミューラー法（Box-Muller transform）により正規分布に従う値を生成する．まず(0,1]の一様乱数をボックス＝ミューラー法で変換し

	A	B	C	D	E	F	G	H	I
1	The number of samples	Discrimination ratio	SubSet1	SubSet2	SubSet3	SubSet4	SubSet5	SubSet6	SubSet7
2	50	0.5	79	69	68	75	67	65	66
3	Sample01		69.464	88.807	85.620	51.781	71.230	67.317	74.557
4	Sample02		43.187	67.496	87.799	67.373	60.458	70.406	44.965
5	Sample03		51.586	87.362	44.421	78.064	87.416	68.212	37.612
6	Sample04		65.715	78.443	72.255	103.001	47.315	74.476	72.467
7	Sample05		86.871	60.704	74.651	70.051	66.301	107.823	54.464
8	Sample06		55.315	73.265	77.057	61.036	91.466	84.714	74.355
9	Sample07		88.681	58.744	87.961	56.976	82.565	74.909	52.282
10	Sample08		55.487	46.125	87.706	85.480	71.783	68.640	85.335
11	Sample09		57.313	64.538	69.991	54.567	59.900	47.345	58.126
12	Sample10		50.578	30.335	67.114	32.162	51.679	63.546	86.093
13	Sample11		55.120	76.098	45.248	58.668	57.561	84.180	111.662
14	Sample12		71.815	64.939	62.306	65.458	83.446	76.021	29.903
15	Sample13		70.299	65.362	92.321	76.867	102.578	79.474	84.803
16	Sample14		56.656	77.004	65.213	71.434	109.265	56.048	48.611
17	Sample15		63.230	125.312	85.455	58.244	108.208	94.949	89.060
18	Sample16		41.459	73.001	60.315	85.852	40.998	107.567	37.584
19	Sample17		28.690	76.106	81.775	110.679	79.097	71.021	48.223
20	Sample18		58.624	70.810	65.998	56.306	117.277	42.678	56.511
21	Sample19		68.617	70.324	61.220	62.056	71.288	63.749	79.045
22	Sample20		61.762	76.354	61.309	59.065	51.231	76.294	95.772
23	Sample21		115.817	57.238	89.282	53.797	77.694	110.687	88.755
24	Sample22		58.674	67.750	80.864	99.398	37.908	59.148	72.692

図２：モデルイメージ

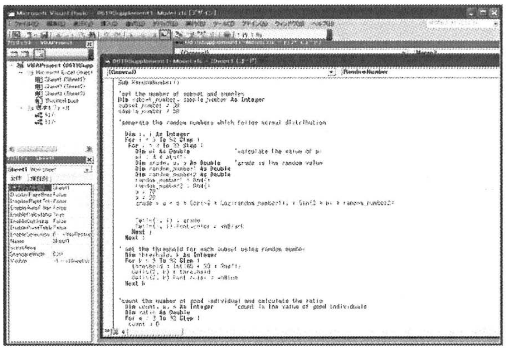

図３：モデルコードの例

て正規乱数をえることから始める．一様乱数(0,1]の要素 a と b を次の変換を用いて変換する.二つの相関のない c と d の正規乱数が次の式にて得られる．

$$c = \sqrt{-2\ln a} * Cos(2\pi b) \tag{1}$$

$$d = \sqrt{-2\ln a} * Sin(2\pi b) \tag{2}$$

本論文では数式（2）の方法を用いて乱数を発生させる．

3.2 サンプリング

各部分集合の生成後，各部分集合に対してしきい値を定める．しきい値は”threshold = Int(60 + 20 * Rnd())”とする．しきい値よりも大きな値の場合は条件を満たしているとする．例えば，sample07 の数は 88.681 であり，しきい値は 79 なので条件を満たしている.そして，全てに対して条件を満たしているかどうか計算し，各部分集合に対しての満たしている比率を求める.もし，比率が判別比よりも大きい場合，部分集合は集合 A に含まれる．そうでなければ，集合 A に含むことはできない．このモデルでは判別比 0.5(50%)としている．

3.3 T-検定

サンプリングの後，サンプルの結果が全体の部分集合の結果と一致するかどうか Student's T 検定により確認を行う．この検定では片側検定を行う．使用する式は以下に記述する．

$$T = \frac{\bar{x} - \mu_0}{\dfrac{S_n^*}{\sqrt{n}}} \tag{3}$$

$\bar{x}$ は各部分集合の平均とする． μ_0 は各部分集合のしきい値とする． S_n^* は各部分集合の修正された標本分散とする．また， n は部分集合の値とする．ここでは有意水準は 5%とする．

検定は以下のように行う．最初に，各部分集合の平均を計算する．次に，修正済の標本分布と T を計算する． λ はエクセルの”TINV”によって得ることができる．T 検定により， $T >= -\lambda$ ならばサンプルの結果を用いる．そして，もし $T <= -\lambda$ ならばサンプリングを棄却し，サンプリングを続けなければならない．

3.4 シミュレーション

このセクションでは，高齢化社会のサンプルに対して，ラフ集合モデルを適用する．

はじめに，ある国に 30 都市があるとする．そして，各都市の 50 人の市民に対してサンプルを行う．全ての世代の市民に対してラフ集合モデルを適用する．モデルにデータ入力をし，データ分類のシミュレーションを行う．しきい値より大きいデータを数え，最後にどの都市が高齢化社会に含むことができるか否か得る．

例えば，50 の都市のデータがあり，各都市では 50 人分の年齢データがある．まず，都市 1 のしきい値を 62 とする．その都市の 62 歳以上の市民の割合が判別比の 0.5(50%)と比較して大きい場合，都市 1 は高齢化社会に入ったと判断する．私たちのモデルに適用し，評価を行う．最初に 62 よりも大きい値を数え，適合度と適合比率が 38，0.76 と計算する．都市 1 の比率が 0.76 と判別比よりはるかに大きいため，第一段階として都市 1 はサンプル結果により高齢化社会に入ったといえる．しかし，その信頼度を確認するために T 検定を用いる．エクセルの結果により T 検定の結果を確認し，サンプリングの結果を受け入れる．そして，都市 1 は高齢化社会に直面していると言うことができる．

都市 2 から都市 30 まで同様に上記のアルゴリズムを繰り返すことにより都市 1 と同様に結果を得ることができる．最後に 30 の都市全ての結果を得ることができる．このセクションでは，サンプルの高齢化社会の状況にラフ集合モデルを適用している．まず，ある国に 30 都市があるとする．各都市の 50 人の市民に対してサンプルを行う．これは全ての世代の市民に対してラフ集合モデルを適用する．モデルにデータ入力をした後，データ分類のシミュレーションを行う．しきい値より大きいデータが数えられ，そして最後にどの都市が高齢化社会に含むことができるか否か得ることができる．

Sample45		79	49	82	65	53	51
Sample46		76	94	78	86	74	40
Sample47		56	61	69	66	56	80
Sample48		89	73	61	71	55	76
Sample49		64	59	88	68	82	81
Sample50		87	67	63	82	55	95
The value of goodness		38	35	25	24	29	18
The ratio of goodness		0.76	0.7	0.5	0.48	0.58	0.36
Is it included in A?		YES	YES	NO	NO	YES	NO
T-Test							
average value of samples		69.61590253	66.88104	69.1652	70.49779	67.77911	67.74423
Sn		401.7886112	510.007	349.7991	401.0574	277.2849	346.3588
T		3.123566407	3.348826	−0.97576	−1.92323	1.588637	−3.48135
λ	1.675905026						
accept or reject?		accept	accept	accept	reject	accept	reject

図 4 ：シミュレーション例

4 終わりに

本研究では現実の様々な問題に対するラフ集合理論を適用した分類方法を提供することを目的とした．また，本論文で提案したモデルではラフ集合を改良した．しかし，現実問題に適用するには多くの問題がある．例えば，現実にデータを集めることは困難である．また，T-検定を行う段階で最初のサンプリングの結果が却下された場合，限りなくサンプリングを行う問題に直面する．したがって，そのような問題を克服するためにラフ集合を適用する更なる研究が必要である．また，ラフ集合を用いて，将来現実問題を解決するために設計を行うべきである．

参考文献

[1] Pawlak, Zdzisław. “Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht” Kluwer Academic Publishing. ISBN 0-7923-1472-7. (1991).

[2] Ziarko, Wojciech. "Rough sets as a methodology for data mining". Rough Sets in Knowledge Discovery 1: Methodology and Applications: 554–576, Heidelberg: Physica-Verlag. (1998).

[3] Ziarko, Wojciech; Shan, Ning. "Discovering attribute relationships, dependencies and rules by using rough sets". Proceedings of the 28th Annual Hawaii International Conference on System Sciences (HICSS'95): 293–299, (1995).

[4] Chen degang, Zhang-wenxiu, “Rough Set and topology spaces, Journal of XianJiaotong University ,1313-1315.(2001).